American Journal of Systems and Software, 2015, Vol. 3, No. 2, 31-35
Available online at http://pubs.sciepub.com/ajss/3/2/1

DOI:10.12691/ajss-3-2-1

A Comparative Analysis of Browsing Behavior of Human Visitors and Automatic Software Agents

Dilip Singh Sisodia[1,*], Shrish Verma[2], Om Prakash Vyas[3]

[1]Department of CS & E, NIT Raipur, Raipur, India
[2]Department of E & TC, NIT Raipur, Raipur, India
[3]Departments of IT, IIIT Allahabad, Allahabad, India
*Corresponding author: dssisodia.cs@nitrr.ac.in

Received February 17, 2015; Revised March 22, 2015; Accepted March 30, 2015

Abstract In this paper, we investigate the comparative access behavior of human visitors and automatic software agents i.e. web robots through access logs of a web portal. We perform an exhaustive investigation on the various resources acquisition trends, hourly activities, entry and exit patterns, geographic analysis of their origin, user agents and the distribution of response sizes and response codes by human visitors and web robots. Gradually web robots are continuing to proliferate and grow in sophistication for non-malicious and malicious reasons. An important share of web traffic is credited to robots and this fraction is likely to cultivate over time. Presence of web robots access traffic entries in web server log repositories imposes a great challenge to extract meaningful knowledge about browsing behavior of actual visitors. This knowledge is useful for enhancement of services for more satisfaction of genuine visitors or optimization of server resources.

Keywords: *software agents web robots, human visitors, resources acquisition, user agents, and response codes*

Cite This Article: Dilip Singh Sisodia, Shrish Verma, and Om Prakash Vyas, "A Comparative Analysis of Browsing Behavior of Human Visitors and Automatic Software Agents." *American Journal of Systems and Software*, vol. 3, no. 2 (2015): 31-35. doi: 10.12691/ajss-3-2-1.

1. Introduction

Largely it is believe that major chunk of web server resources are used to handle human visitor's generated traffic for any web portal. This belief seems to be a myth if we look at recent reports [1] which states that major portion of web traffic is generated through automatic software agents. Most website owners simple rely on web analytics tools to track who's visiting their site, but these tools doesn't show you 51% of your site's traffic including some seriously shady non-human visitors like web robots (also known as web crawlers, spiders, wanderers, and harvesters etc) which is used for non-malicious (by search engines) and malicious reasons (by hackers, scrapers, spammers and spies) of all sorts who are easily thwarted, but only if they're seen and blocked [1]. With the advent of web 2.0 services web robots are the Internet's silent majority behind the scenes, billions of these software agents are hard at work, shaping our web experience and playing a key role in everything we do online. These types of robots include Indexers (or search engine crawlers) seeks to harvest as much web content as possible on a regular basis, in order to build and maintain large search indexes. Analyzers (or shopping bots) crawl the web to compare prices and products sold by different e-Commerce sites. Experimental (focused crawlers) seek and acquire web-pages belonging to pre-specified thematic areas. Harvesters (email harvesters) collect email addresses on behalf of email marketing companies or spammers. Verifier (site-specific crawlers) performs various website maintenance chores, such as mirroring web sites or discovering their broken links. RSS crawler use to retrieve information from RSS feeds of a web site or a blog. Scrapers used to automatically create copies of web sites for malicious purposes. Since their inception (First web robots were introduced in 1993) they are increasing exponentially. Because they are very simple to create and they offer great job by circumvent collection of information [2,3].

This paper is organized as follows: in section 2, we discuss the previous works on access behaviour of robots and human visitors. In section3, we outline the brief description of methodology adopted for this work. In section 4, we present the various comparison metrics and perform the experiments. In section 5, the paper is concluded with final observations.

2. Related Work

The human visitors induced web access behaviour is thoroughly studied by various studies. In [4] series of metrics to describe the aggregate web traffic is proposed, the self-similarity in Web access pattern had been discussed in [5]. A particular web site's visitors are classified into different groups according to their purchase habits in [6]. In [7] Cluster based analysis to classify large number of sessions into several coherent classes that

efficiently describe web server workloads. This inclusive study of human visitors induced web traffic leads to design for various solutions like effective and efficient web site design [8] and optimal cache replacement policies [9,10] according to the variations in human navigation patterns [11,12].Web robots access behaviour had characterize in [13] by using some popular robots. In [14] an empirical study was carried out on a very large data to classify various robots on the basis of their access behaviour. In [15] robots behaviour was studied in scholarly information environment. In [16] authors present the inherent access pattern of crawlers on online social network sites (OSNS). But most of these findings were done either before emergence of web 2.0 based services or concentrated on web robots or human visitors in isolated way. We did not find any significant amount of literature except [17] which compare and differentiate between the web access behaviour of web robots and human visitors. In [17] authors compared only resource request patterns of both type of visitors and ignores the other aspects of web access behaviour of web robots and human visitors. Here our major contribution will be an exhaustive investigation on the hourly activities, entry and exit patterns, various resources acquisition trends, geographic analysis of their origin, user agents and the distribution of response sizes and response codes by human visitors and web robots.

3. Methodology

In this section we going to present a brief description of methodology adopted for this work through flowchart (Figure 1). For this analysis we are considering web server access log of a portal which provides assistance for visa, insurance and other related services to worldwide visitors in USA.

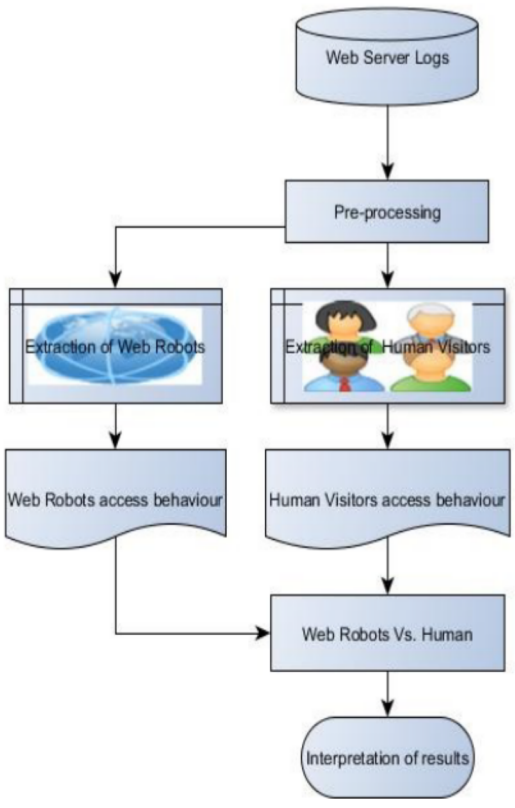

Figure 1. Methodology for comparison of of Web Robots vs. Humans

3.1. Data Set Description

Web server access logs: Web log file consists of four kinds of records: access log, error log, referrer log and agent log. The forms of web log file are usually of two types, common log file (CLF) and extended log file (ELF). A web server access log store the detailed information of each request made from user's web browsers to the web server in a chronological order [18]. An example of classic web server log entries is given as follows and brief description in Table 1.

11.111.11.111 - - [15/Dec/2013:00:01:02 -0800] "GET/forum/member.php?45067-Carla-Zenis&tab=activitystream&type=all HTTP/1.1"200 10463"http://www.google.com/bot.html" "Mozilla/5.0 (compatible; Googlebot/2.1)"

Table 1. Brief description of web log entry headers

Log entry headers	Description of headers
11.111.11.111	Remote host address
-	Remote log name
-	User name
15/Dec/2013:00:01:02	Timestamp
-800	Time zone of the request
GET	Request method
(/forum/member.php?45067-Carla-Zenis&tab=activitystream&type=all	Path on the server
HTTP/1.1	Protocol version
200	Service status code
10463	Size of the returned data
http://www.google.com/bot.html	*Referrer
Mozilla/5.0 (compatible; Googlebot/2.1)	*User agent

*These fields only appear in the extended log file (ELF).

3.2. Pre-Processing

This log has parsed and pre-processed with the help of customized program which is based on an open source tool [19]. We extract various statistical log information like number of requests, request duration, number of users, page hits, domains and countries of host visitors, used operating system, used browser, robots activity, HTTP errors and many more from given log.

3.3. Identification of Robot and Human Sessions

Web robot sessions are extracted by using multi fold approaches in first step we applied well known heuristics proposed in [2] but this has left number of robot sessions without identification. So in second step our log analyzer uses the database of IP addresses and user agent fields of well known bots [20,21]. if the web serve log session's IP addresses or user agent is matches with IP or user agent of well known crawlers then session is labelled as web robot sessions and it effectively obtains a sizable sample of requests from web robots to infer significant trends. Human visitor's sessions are identified by using time oriented heuristics. The time-oriented heuristic can be two types: the session-duration heuristic and page-stay time heuristic. These two time-oriented heuristics can be used individually or jointly to segment the user activity log into sessions. Each heuristic h scans the user activity logs to which the web server log is partitioned after user identification, and outputs a set of constructed sessions:

• h1 (session-duration heuristic): Total session duration may not exceed a threshold θ. Given t0, the timestamp for

the first request in a constructed session S, the request with a timestamp t is assigned to S, iff $t - t_0 \leq \theta$. This heuristics varies from 25.5 minutes to 24 hours while 30 minutes is the mostly used default timeout for session duration [22].This information is summarized in Table 2.

Table 2. web server access log of twenty day's duration

Parameters	Total Traffic		Human induced traffic		Web robots induced traffic	
	Before Pre-processing	After Pre-processing	Before Pre-processing	After Pre-processing	Before Pre-processing	After Pre-processing
# of requests	14022101	4432393	13427825	3942093	594276	490300
#Avg.req/day	701105	221619	671391	197104	29713	24515
# of users	213864	182337	210694	179814	3170	2636
# of sessions	471974	319554	447317	303826	24657	15728
Bandwidth	93.9 GB	28.71 GB	88.73 GB	24.56GB	5.17 GB	4.16 GB

4. Experiments

In this section we perform various experiments to draw the comparison between access behaviour of human visitors and web robots. As shown in Table 2, very large number of requests, users and sessions are generated for the web server logs. To capture microscopic view and avoid processing overhead for this analysis we are using sample of this log to show comparisons.

4.1. Experiment-I

Comparison of Resource acquisition pattern:- Here we analyze and compare the percentage of requests, percentage of visitors and percentage of bandwidth consumed by human and robots for each specific types of resources.. The most striking observation to emerge from the data comparison (Figure 2) was that robots exhibit their aggression to only access web resources (*.html, *.php, *.htm etc.) and engrossed more number of requests, visitors and consumed more bandwidth as compare to human visitors. While Human visitors show uniform access behaviour for all type of resources and receive less number of requests and visitors and consumed less bandwidth for web resources but aggregate value (including all type resources) is quite high as compared to robots. Interestingly, there were also differences in the ratios of web to image resources accessed by humans and robots. This value is very large for robots than humans. It is reasonable to expect humans to request many web resources as they browse from page to page to retrieve information and download files. But this percentage may be low because humans' liking may be twisted towards embedded resources (*.jpg, *.png,*.gif etc.) with web pages.

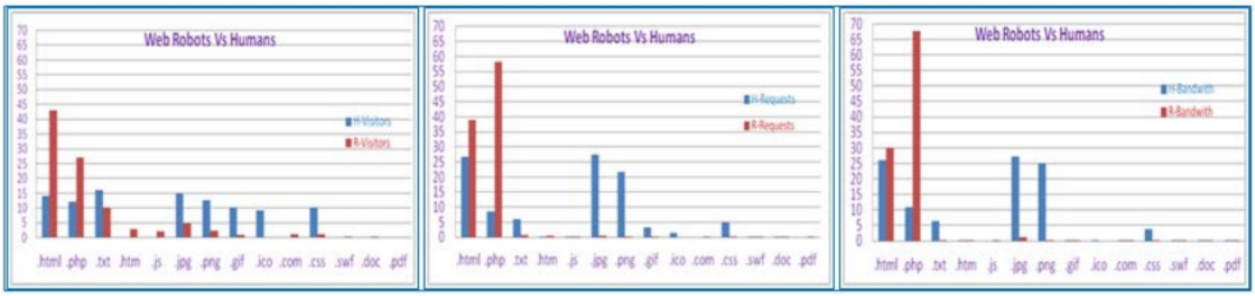

Figure 2. % of Requests Received, % of Visitors Attracted and % of Bandwidth consumed by different types of resources

Figure 3. Comparison access behavior of Web robots Vs Humans

4.2. Experiment-II

Comparison of general browsing behaviour: - In this experiment, we examine the hourly distributions of % of requests, % of visitors and % of bandwidth consumed by robots and humans.

We also examined the access behaviour of robots and humans for most popular resources, top entry pages and top exit pages. It is evident from the experiment (Figure 3) that the human visitors exhibit a consistent access tendency throughout the day but robots initially (red spikes) generates vast amount of traffic to request large number of resources and consumed significant amount of band width. Popular resources are accessed by robots and humans are localized in different localities but human visitors are monotonous and restricted to few web resources while robots perform exhaustive search for multiplicity of resources and followed diverse entry and exit paths.3.

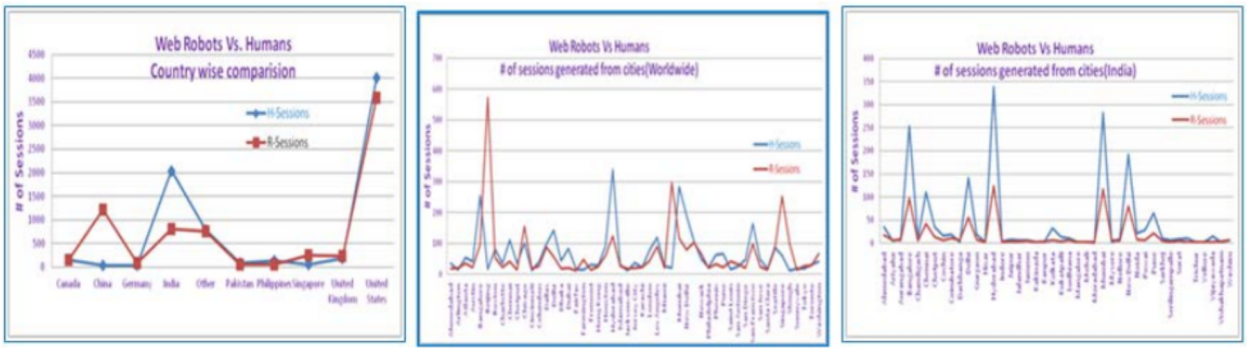

Figure 4. Comparison of demographic browsing behaviour of Web robots Vs Humans

4.3. Experiment-III

Comparison of Demographic browsing behaviour: - In this analysis we are examining the demographic origin of robots and human visitors (Figure 4). Largest share of visitors sessions (both robots and humans) are credited to USA and followed by India. China is the only country who had made significant contribution to web robots sessions but very small share of human visitors. Among global cites most of the robot sessions are originated from Beijing, Mountain View, Chicago and Singapore etc. while human sessions are generated across the world but major share contributed from Indian cities (Hyderabad, Bangalore, Mumbai etc.) where Hi-tech industries are blooming, USA cities (San Francisco, Loss Angeles, Chicago etc.), Singapore and Dubai. In India both robots and human sessions are generated from almost same cities but quantum of human sessions is much more than robot sessions.

4.4. Experiment-IV

Comparison of Access paths and Response codes:- Here we will discuss the path followed, responses received and operating systems used by web robots and human visitors (Figure 5). Human visitors followed diverse paths and most of the time received successful response from the sever as compared to robots.web robots used very long paths and generates high frequency for these paths because they are mechanised to do the same job repeatedly.web robots also receive different types of erroneous responses from servers because they are automatic software agents may follow the broken links on web pages or try to get resources which are not available. Both human and robots used different types of OS but humans are dominated in GUI based desktop OS systems while robots used server OS along with GUI based desktop OS.

Figure 5. Comparison of Access paths, Response codes and Operating Systems

5. Conclusions and Future Work

The main goal of this investigation was to determine comparative browsing behaviour of the human visitors and Web robots. This analysis provided insights into the number of requests generated, type of resources retrieved, amount of bandwidth consumed by web robots and human visitors. We also exposed similarities and differences between their access paths, responses received from servers and their preferred entry and exit pages. The empirical findings in this study provide a new understanding for demographic origination of humans and robots. The most obvious finding to emerge from this study is that web robot traffic may cause the performance

bottleneck of web server because it drains significant portion of the server resources if it remains unnoticed. Future work will use these findings to develop new techniques for identification of web robots and minimize their impact on server resources. Some more logs from numerous domains will also be analyzed to further confirm our findings.

References

[1] http://www.incapsula.com/blog/what-google-doesnt-show-you-31-of-website-traffic-can-harm-your-business.html.

[2] P. N. Tan and V. Kumar, "Discovery of web robot sessions based on their navigational patterns," Data Mining and Knowledge Discovery, vol. 6, pp. 9-35, 2002.

[3] D. Doran and S. S. Gokhale. Web Robot Detection Techniques: Overview and Limitations. Data Mining and Knowledge Discovery, 22(1-2):183-210, 2011.

[4] M. F. Arlitt and C. L. Williamson, "Web server workload characterization: The search for invariants," ACM SIGMETRICS Performance Evaluation Review, pp. 126-137, 1996.

[5] Mark E. Crovella and Azer Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. Transactions on Networking, 5(6):835-846, December 1997.

[6] J. X. Yu, Y. Ou, C. Zhang, and S. Zhang, "Identifying interesting customers through web log classification," IEEE Intelligent Systems, vol. 20, no. 3, pp. 55-59, 2005.

[7] F. Li, K. Goseva-Popstojanova, and A. Ross, "Discovering web workload characteristics through cluster analysis," in Proc. IEEE International Symposium on Network Computing and Applications, 2007, pp. 61-68.

[8] M. Spiliopoulou, "Web usage mining for web site evaluation," Communications of the ACM, vol. 43, no. 8, 2000.

[9] M.-L. Shyu, C. Haruechaiyasak, and S.-C. Chen, "Mining user access patterns with traversal constraint for predicting web page requests," Knowl. Inf. Syst., vol. 10, no. 4, pp. 515-528, 2006.

[10] Almeida, V., Menascé, D., Riedi, R., Peligrinelli, F., Fonseca, R., & Meira Jr, W. (2001, June). Analyzing Web robots and their impact on caching. In Proc. Sixth Workshop on Web Caching and Content Distribution (pp. 20-22).

[11] R. White and S. Drucker, "Investigating behavioral variability in web search," in Proc. of the 16th Intl. conference on World Wide Web. ACM, 2007, pp. 21-30.

[12] X. Lin, L. Quan, and H. Wu, "An automatic scheme to categorize user sessions in modern http traffic," in Proc. Of IEEE Global Telecommunications Conference (GLOBECOM 08), New Orleans, LO, November 2008, pp. 1-6.

[13] M. D. Dikaiakosa, A. Stassopouloub, and L. Papageorgioua. An Investigation of Web Crawler Behavior: Characterization and Metrics. Computer Networks, 28:880-897, 2005.

[14] Lee, Junsup, Sungdeok Cha, Dongkun Lee, and Hyungkyu Lee. "Classification of web robots: An empirical study based on over one billion requests." computers & security 28, no. 8 (2009): 795-802.

[15] P. Huntington, D. Nicholas, and H. R. Jamali, "Web robot detection in the scholarly information environment," Journal of Information Science, vol. 34, no. 5, pp. 726-741, 2008.

[16] G. Jacob, E. Kirda, C. Kruegel, and G. Vigna, "PUBCRAWL: protecting users and businesses from crawlers," in Proceedings of the 21st USENIX conference on Security symposium. USENIX Association, 2012.

[17] Doran, Derek, Kevin Morillo, and Swapna S. Gokhale. "A comparison of web robot and human requests." In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1374-1380. ACM, 2013.

[18] Sisodia, Dilip Singh, and Shrish Verma. "Web usage pattern analysis through web logs: A review." In Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on, pp. 49-53. IEEE, 2012.

[19] "AWStats - free log file analyzer for advanced statistics (GNU GPL), http://awstats.sourceforge.net/.(accesed in February 2014)

[20] User agents database http://www.user-agents.org/index.shtml (accessed in February 2014).

[21] Well known robots database http://www.robotstxt.org/db.html(accesed in February 2014).

[22] Berendt, B., Mobasher, B., Spiliopoulou, M., and Nakagawa, M. "A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis," INFORMS Journal of Computing, Special Issue on Mining Web-Based Data for E-Business Applications Vol. 15, No. 2, 2003.